June 19, 2017

Attn: William Hoang

Springboard

New York, NY

**Re: Capstone Project Proposal**

**Introduction**:

Walmart’s acquisition of Jet.com for $1.3 billion and Amazon’s acquisition of Whole Foods for $13.7 billion are symptomatic of an anxiety surrounding the future of retail. No superlatives have been spared to describe the trend, despite eCommerce accounting for no more than 9% of US retail spend, as of Q1 2017 (according to the Census Bureau). At its very core, this shift falls under the domain of Supply Chain Management; products and services are experiencing a wildly different lifecycle between the point of production and the point of consumption.

The result is that retailers with lean, adaptive supply chains will win. The implicit imperative behind retail’s changing landscape is the need to predict and adapt to consumer habits, which makes for some supremely interesting Supply Chain questions.

**Instacart Kaggle Competition**:

It wasn’t accidental that I mentioned Whole Foods above. Instacart is an online grocery and delivery app. Instacart’s shopping experience is driven by data scientists that curate the products that you see. The current $20 billion Americans spend at online groceries represents a meager 2.5% of the estimated $800 billion US grocery market, where perishable goods make predicting consumer buying habits a matter of life or death (figuratively, except for the produce). Demand Forecasting being a cornerstone of Supply Chain Management (SCM), I see this challenge as an opportunity to explore the domains that have driven my career thus far – eCommerce and SCM – through the lens of data science.

Please consult the competition website for more details:

https://www.kaggle.com/c/instacart-market-basket-analysis

**Dataset Overview**:

Instacart provides an open-source dataset of over 3 million orders from 200,000 users, in a competition aimed at producing an algorithm capable of predicting what returning users will order next. Instacart will use the winning submission to refine its product recommendations and personalize the digital shopping experience.

The dataset includes order data, like basket assortment, day of the week, and hour of the day, as well as anonymized customer and product data, like order history and product category. Submissions must be a list of products corresponding to a given order, and will be evaluated using mean F1 score.

**Client**:

Instacart.

**Who Cares?**

For starters, Instacart, which is shelling out $25,000 for this competition and accepting resumes for top performers. Broadly speaking, predictive algorithms in the Demand Forecasting space are hugely valuable in the evolving realm of retail eCommerce, where logistics present unique difficulties for perishable and frozen products, among others, that generally have never been put in a USPS box (until recently).

The ability to forecast demand, even if the resulting efficacy is ever-so-slight, will have a giant impact on Instacart’s supply chain and logistics operations, let alone their customers, who will benefit from more personalized experiences, fresher groceries, and foregone queues at the market. In an unnecessary summation of the value of predictive algorithms: one day means a lot to a banana (microphone dropped).

(Microphone retrieved) While a specific online grocer will immediately benefit from this analysis, demand forecasting in eCommerce is a critical part of SCM, and having run operations of an eCommerce home improvement retailer for the past three years, I can say with certainty that I could have made use of the resources and skills that I’ll benefit from over the course of this project.

**Methodology**:

The Instacart team has provided substantial context for the dataset on their blog (e.g. ice-cream products represent 24 of the 25 top products bought late at night):

<https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2>

One of the most compelling aspects of this competition vis-à-vis Springboard’s Capstone Project is that the data has already been wrangled and is relatively clean, which allows me to jump into Exploratory Data Analysis. I intend to use pandas and numpy for analytics, and matplotlib for visualization (potentially Seaborn or Bokeh – I’d like to learn Bokeh in this process). I will also look into XGBoost (which Instacart uses in production) and Keras for machine learning. I will likely make use of scikit-learn for machine learning also.

My immediate goal is to substantially increase my skills in inferential statistics, data visualization, and predictive modelling, as well as becoming conversant with the fundamentals of machine learning techniques. These areas will be the focus of my approach.

My ultimate goal is to use this experience, including immersing myself in the Kaggle discussions and culture of collaborative open-source analysis, to be able to function at a high level within an organization both from an Operations & Strategy perspective, as well as from a Data Science perspective, focusing on applied strategy and decision-making.

**Deliverables**:

I will most likely deliver the code in the form of a Jupyter notebook documenting the Exploratory Data Analysis and resulting algorithm, in addition to the competition submission itself. The end product, then, is a model capable of reproducing the predictions in the submission. I will also produce a formal thesis on my findings, as well as a deck for the faint-hearted.

I’m looking forward to jumping in. Thank you very much!

Respectfully,

Ryan Alexander Alberts